Cell Reports, Volume *34*

Supplemental information

Most non-canonical proteins uniquely populate the proteome or immunopeptidome

Maria Virginia Ruiz Cuevas, Marie-Pierre Hardy, Jaroslav Hollý, Éric Bonneil, Chantal Durette, Mathieu Courcelles, Joël Lanoix, Caroline Côté, Louis M. Staudt, Sébastien Lemieux, Pierre Thibault, Claude Perreault, and Jonathan W. Yewdell

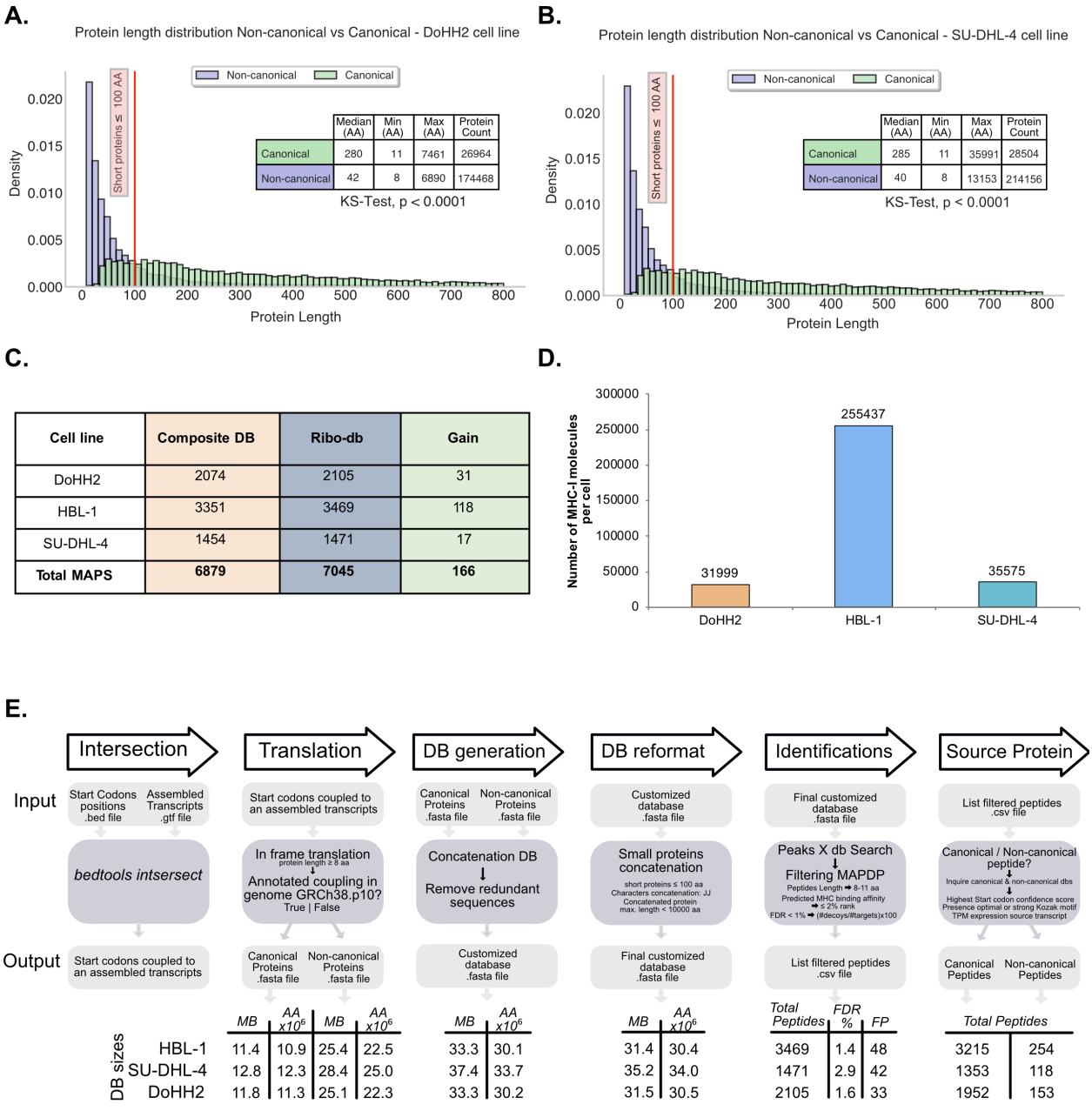

Figure S1. Related to Figure 1. Sample-specific database composition.
(A) Length distribution of canonical and non-canonical proteins from DoHH2 database showed significant differences. P< 0.0001, Kolmogorov-Smirnov Test. Total proteins for both categories are indicated on the legend, besides their median, minimum (Min) and maximum (Max) observed lengths. Proteins with a length >800 AA are not displayed on the graph.
(B) Length distribution of canonical and non-canonical proteins from SU-DHL-4 database showed significant differences. P< 0.0001, Kolmogorov-Smirnov Test. Total proteins for both categories are indicated on the table legend, besides their median, minimum (Min) and maximum (Max) observed lengths. Proteins with a length >800 AA are not displayed on the graph.
(C) MAPs identified through composite databases Ribo-db + PRICE method vs MAPs identified solely on Ribo-db-derived databases. MS-Peaks database searches performed solely on Ribo-db allowed to gain (2%) more MAPs dentifications.
(D) Absolute number of MHC-I molecules per cell, in 3 cell lines, measured by flow cytometry using QIFIKIT (see Methods).
(E) MAPs identification process. Diagram that details, from the intersect step of the general workflow overview (Figure 1A), the strategy used for database generation and the filtering for the MAPs identification. The size (Mb) and the total number of amino acids (AA) of each database are shown at the bottom of the figure, along with the FDR used and the number of expected erroneous identifications (False Positive, FP).

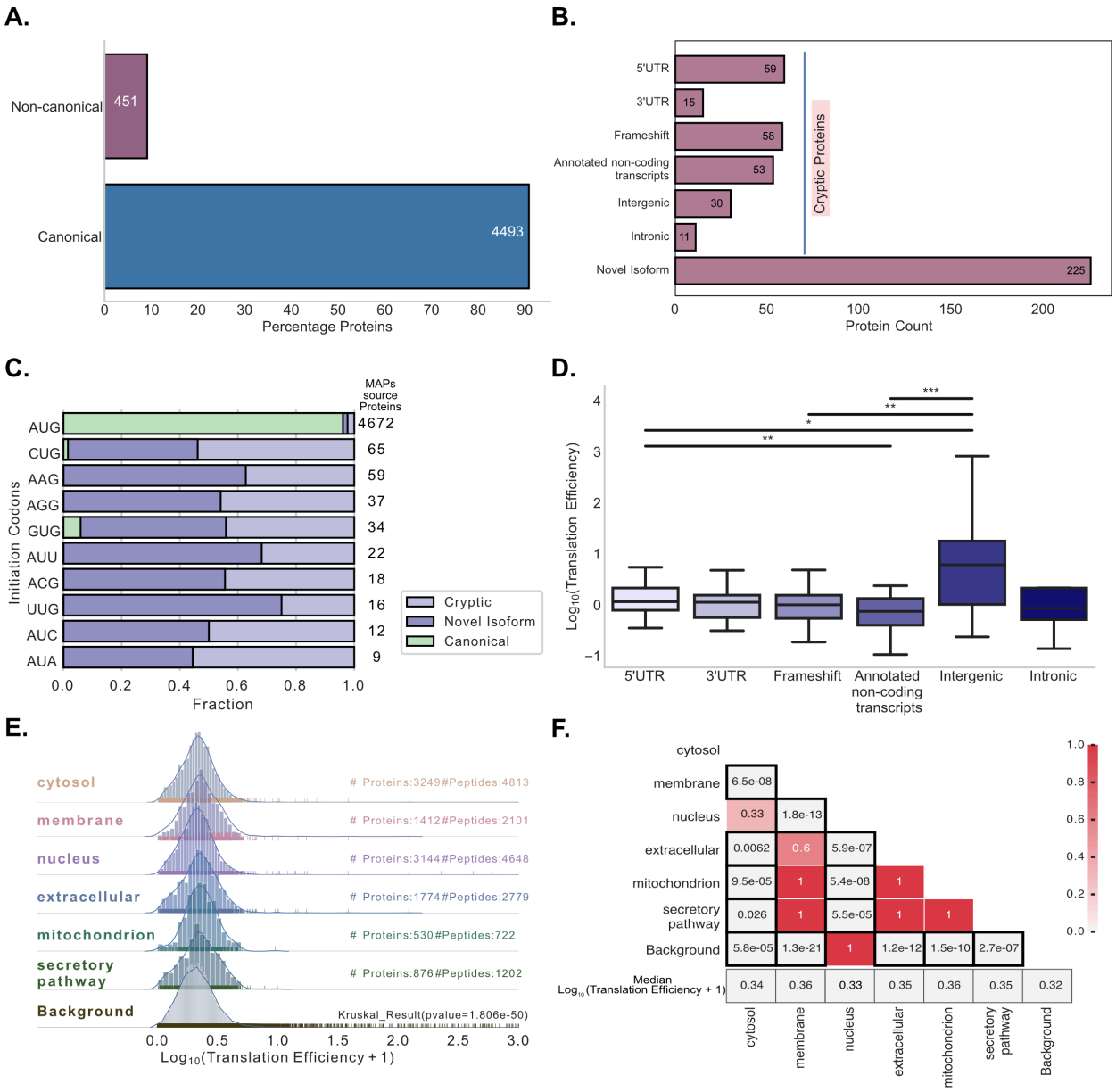

Figure S2. Related to Figures 2,3. Properties of the novel proteins identified in the immunopeptidome analysis.

(A) At least 10% of the MAPs source proteins derived from non-canonical proteins. Bar plot depicting the percentage of proteins source of MAPs. The purple bar shows the percentage of cryptic and novel isoform proteins (10%), the blue bar shows the percentage of canonical proteins (90%).

(B) Protein count of the MAPs cryptic and novel isoform proteins. The bar plot depicts the total number for each category of the Cryptic along with the total number of Novel Isoform proteins.

(C) Most of the new proteins initiated at near cognate codons. Bar plots showing the fraction of cryptic, novel isoform and canonical proteins initiated through each initiation codon.

(D) Translation efficiency of the MAPs source cryptic proteins. Boxplots showing the translation efficiency distribution for each one of the categories into the MAPs source cryptic proteins. Translational efficiency of each MAPs source protein was calculated as the ratio of translation (derived from counts of ribosome profiling reads) to transcription (derived from RNA-seq reads). *P< 0.05, **P< 0.01, ***P< 0.001, ****P< 0.0001, two-side Mann-Whitney U Test corrected with Bonferroni correction.

(E) Translation efficiency distributions of MAPs source proteins according to their subcellular localization. Background proteins are canonical proteins non-source of MAPs. Number of proteins and number of peptides are presented for each localization. Statistical difference was assessed by Kruskal-Wallis.

(F) Heatmap presenting the adjusted p-values according to the post-hoc comparison for the translation efficiency of the canonical MAPs source proteins. Statistical differences were assessed by Mann–Whitney U tests with Bonferroni correction.

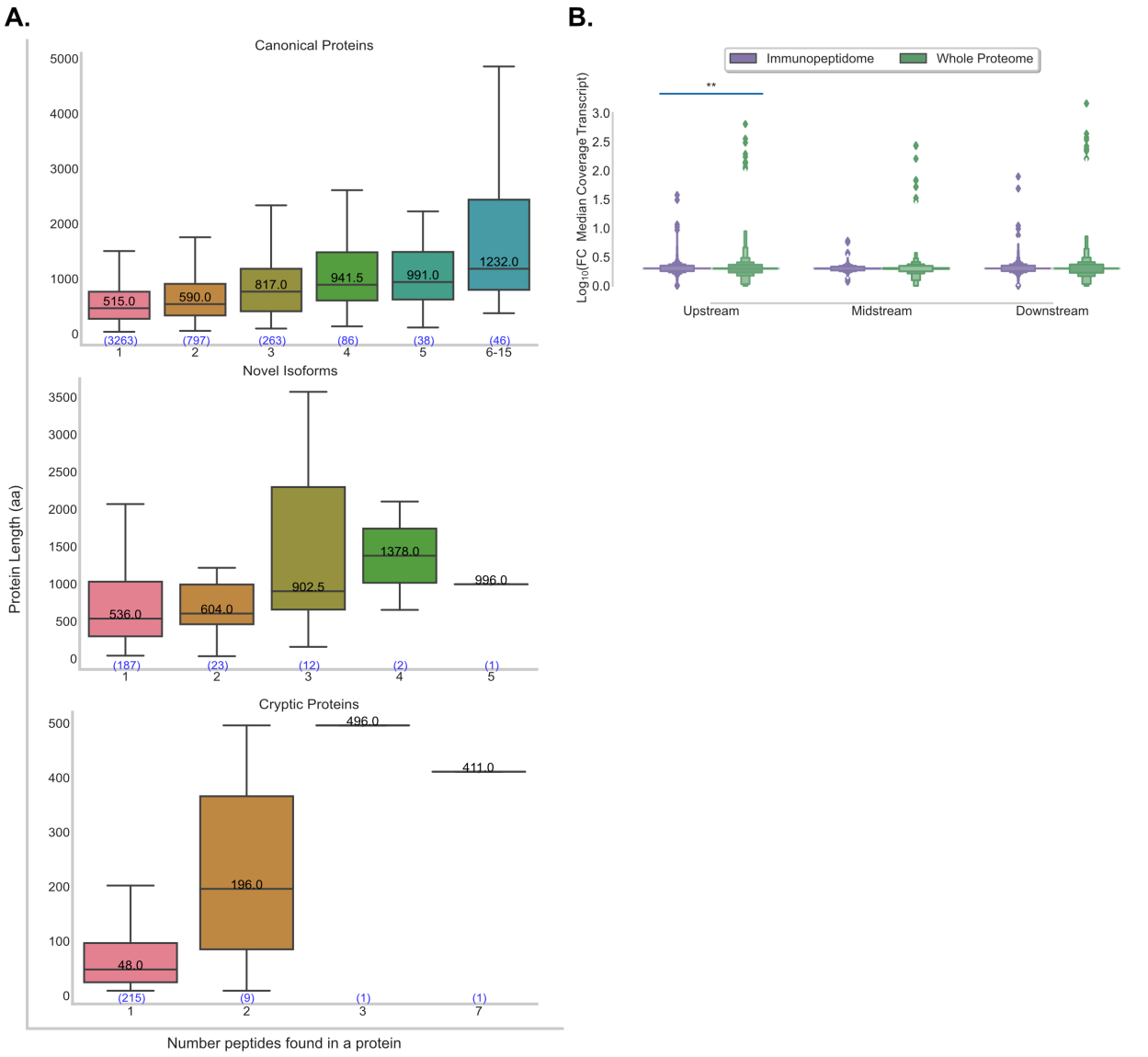

Figure S3. Related to Figures 3 and 5. Properties of the novel proteins identified in the immunopeptidome analysis.
(A) Boxplot graphs for canonical proteins, novel isoforms and cryptic proteins, showing the number of identified MAPs vs. length of the source protein. The median length of the proteins is shown into each boxplot. The number of proteins that have the number of peptides specified on the X-axis is shown at the bottom of the box chart (blue numbers).
(B) Boxplots showing the fold change of the median coverage for Up-Mid-Downstream relative to the median coverage of the whole transcript, for the MAPs source proteins vs the non-source proteins detected in the whole proteome analysis. The resulting distributions are plotted in log10 of the fold change. Statistical difference was assessed by Wilcoxon signed-rank test.

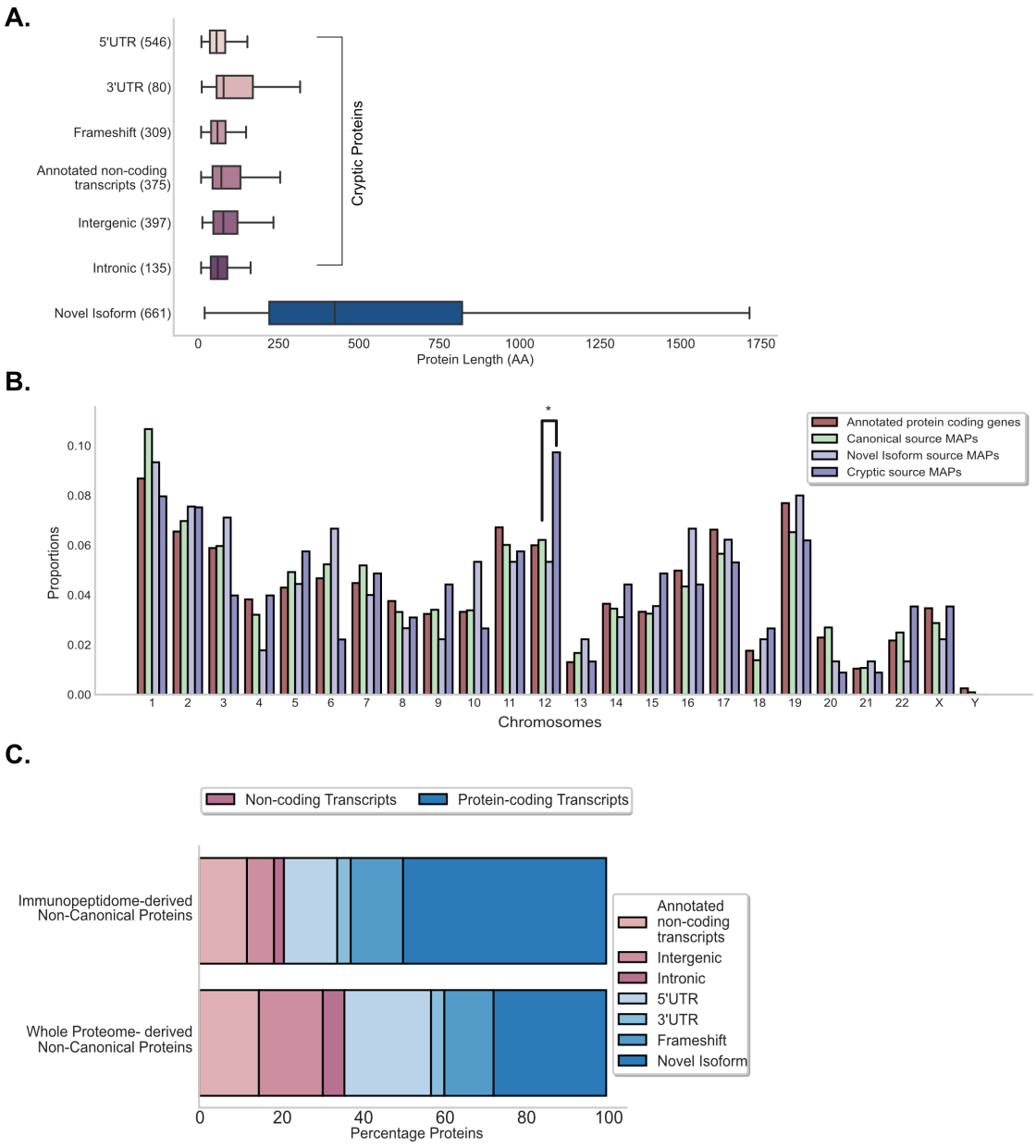

Figure S4. Related to Figure 6. Features of the newly elucidated proteins.
(A) Cryptic proteins are significantly shorter than novel isoform proteins. Boxplots indicating the length distribution of the newly identified proteins for each category: cryptic proteins (5'UTR, 3'UTR, frameshift, annotated noncoding transcripts, intergenic, intronic) and novel isoforms. Median length in novel isoform (448 aa) and cryptic proteins (65 aa) differed significantly according to two-side Mann-Whitney U test, ****P< 0.0001.
(B) MAPs source newly identified proteins derive from all chromosomes. Bar graph showing, in proportion, the chromosomal origin of each category of proteins, compared to canonical protein coding genes. Chromosome 12 appeared to be rich in cryptic proteins, *P< 0.05, two-side Fisher's exact test.
(C) MAPs source novel proteins derived preferentially from protein-coding transcripts (79% protein-coding vs 21% noncoding transcripts) compared to the percentages on the whole non-canonical proteome (whole proteome analysis-derived proteins). Stacked bar plot showing the percentage of novel identified proteins deriving either noncoding (red bars) and protein-coding transcripts (blue bars) for MAPs non-canonical source proteins vs whole non-canonical proteome.

Table S1. Related to Figure 1C. Size of protein databases used in this study.

	Database size (MB)		
	PRICE	**Ribo-db**	**Composite db Ribo-db +PRICE**
DoHH2	*10.4*	*31.5*	*41.9*
SU-DHL-4	*10.5*	*35.2*	*45.7*
HBL-1	*11.8*	*31.4*	*43.2*

Table S2. Related to Methods: Ribo-db approach: detection of active translation sequences. Number of MS identified proteins in the 3 cell lines.

			MS-Identified Proteins		
		Proteins in the DB	**Immunopeptidome**	**Whole Proteome***	**Total**
DoHH2	*Canonical*	26964	1366	2478	4612
	Non-canonical	174468	141	627	
SU-DHL-4	*Canonical*	28504	1017	2798	4712
	Non-canonical	214156	95	802	
HBL-1	*Canonical*	26726	2110	2226	5174
	Non-canonical	181811	215	623	
		Total	4944	9554	14498

*(without overlapped proteins in Immunopeptidome)

Table S3. Related to Methods: Ribo-db approach: detection of active translation sequences. Percentages of MS-detected proteins among Ribo-seq identified proteins.

	% (MS / total db)	
	Immunopeptidome	**Proteome**
Canonical	5.51	9.11
Non-canonical	0.08	0.36